Bellaterra Journal of Teaching & Learning Language & Literature
Vol. 10(4), Nov-Dec 2017, 68-86

Investigating the use of readability metrics to detect differences in written productions of learners: a corpus-based study

Paula Lissón
Université Paris Diderot (USPC)

(Text received 21 December 2017; accepted 31 December 2017)
DOI: https://doi.org/10.5565/rev/jtl3.752

Abstract: This paper deals with the use of readability metrics as indices of learners' linguistic features in a written corpus of Spanish learners of English L2. Seventeen measures of readability are presented and computed for 200 samples of written argumentative essays extracted from the corpus NOCE (Díaz-Negrillo, 2007). Support Vector Machines (SVM) are used in order to detect which are the metrics that perform better at detecting differences in learners' productions belonging to students enrolled in the first or in the second year of an English major. Metrics based on sentence length, number of sentences, and number of polysyllabic words are reported to be the most accurate ones for the classification of learners' linguistic features.

Keywords: readability, learner corpora, SVM, written essays

Résumé: Cet article analyse les métriques de lisibilité dans une perspective de détection des niveaux des apprenants hispanophones en anglais L2 à partir d'un corpus de productions écrites. Dix-sept métriques de lisibilité ont été calculées pour 200 essais argumentatifs extraits du corpus NOCE (Díaz-Negrillo, 2007). On emploie la technique des Séparateurs à Vaste Marge (SVM) pour déterminer les métriques qui caractérisent le mieux les différences entre les productions des apprenants de première et deuxième année en anglais. Les métriques fondées sur la longueur des phrases, le nombre de phrases, ainsi que sur le nombre de mots polysyllabiques sont celles qui aident le mieux à détecter et à classer les étudiants selon leurs caractéristiques langagières.

Mots-clés: lisibilité, corpus d'apprenants, SVM, productions écrites

Resumen: Este artículo trata sobre el uso de métricas de legibilidad como indicadores de las características lingüísticas propias a dos niveles de aprendices españoles de inglés L2. Presentamos y calculamos diecisiete medidas de legibilidad en 200 textos argumentativos extraídos del corpus NOCE (Díaz-Negrillo, 2007). Utilizamos SVM para averiguar qué métricas son capaces de detectar diferencias entre las 200 producciones, pertenecientes a alumnos de primer y segundo curso de Filología Inglesa, respectivamente. Las métricas basadas en la longitud de las frases, el número de frases y el número de palabras polisílabas son las que presentan mejores resultados.

Palabras clave: legibilidad, corpus de aprendices, SVM, producciones escritas

Introduction

This paper is an empirical study that deals with the use of readability metrics as indices of linguistic features when comparing two groups of learners of English. The goal of this study is to show that readability metrics, combined with a machine learning approach, can be used to detect clusters of writing features that correspond to different groups of learners. The introduction of this paper is a brief overview on the concept of readability and the use of readability metrics in Second Language Acquisition (SLA) and Learner Corpus Research (LCR). The rest of the paper is organized as follows: section 2 describes the corpus and the readability metrics used in this study; section 3 presents raw results and analysis of the scores; section 4 introduces the use of Support Vector Machines (SVM) in linguistics and Natural Language Processing (NLP), as well as the results for the data under scrutiny; finally, section 5 presents main conclusions and a discussion.

Readability and readability metrics in SLA and LCR. Although the difficulty of a text may be measured in many ways and according to several features, readability refers to the understanding of the text as such, i.e., how easy or how difficult a text is to be understood by a reader or a group of readers. One of the most famous definitions, written by Dale and Chall (as cited in DuBay, 2004, p. 3) states that readability is composed by "[…] all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting". McLaughlin (1968, p. 188) creator of the SMOG readability formula, considers *readability* to be "the degree to which a given class of people find certain reading matter compelling, and necessarily, comprehensible", and according to DuBay (2004, p.3), readability is simply "what makes some texts easier to read than others". Therefore, readability can be seen as the combination of linguistic features that make a text more or less readable for a particular audience. In that sense, the aim of the creation of readability metrics is to account for text readability, somehow measuring or quantifying the difficulty or the easiness of written materials.

Readability metrics are essentially mathematical formulae containing different linguistic features that "map a text to a numerical value corresponding to a difficulty or grade level" (Heilman, Collins-Thompson, & Eskenazi, 2008, p. 71). Most of the classical readability metrics are built on linear models (Sung, et al., 2015, p. 340), and parameters of the formulae are usually related to lexical and syntactic features, such as the use of content vs. function words, lexical diversity, or syllable counts. Recent approaches to readability metrics,

however, include more complex algorithms with a larger number of parameters (see, for example, François, 2011), some of them related to unigrams modelling (Collins-Thompson & Callan, 2005), or discourse-based characteristics (Feng, Jansche, Huenerfauth, & Elhadad, 2010), among many other possibilities.

Traditionally, the role of readability measures in SLA has been connected to the difficulty of the texts that were to be read by learners. In other words, readability measures have been used in order to determine if a text is appropriate or not for learners of a particular level. In that sense, the teaching applications of readability measures are straightforward: teachers may need to know if the texts they are using are adequate for their students. Thus, readability metrics have been generally applied to texts that could (or could not) be used for learners of a certain level or of a particular group age. Readability formulae were also typically applied to the comparison of textbooks, in order to assess the complexity of the manuals. Nowadays, however, with the emergence of the field of Learner Corpus Research (see Granger, Gilquin, & Meunier, 2015) and the massive collection of learners' productions, new ways of applying readability to the study of SLA and language teaching have appeared. For instance, Vajjala and Meurers (2012) showed that application of learners' SLA based lexical features as part of the readability formulae yielded high accuracy in readability results. Lu (2010) created an automatic system for the assessment of learners' productions based on some of the parameters used in most readability metrics.

In this paper, I aim to change the traditional point of view of readability metrics. Following Lu (2011) and Vajjala (2018), I am not using readability metrics in order to see how difficult a text might be for a given level of proficiency. Rather, I am applying readability formulae to learners' productions so as to see what readability metrics can tell us about learners' use of linguistic features and the differences between two groups of learners. This approach is based on the claim that there exist specific linguistic features, also known as 'criterial features' (Hawkins & Buttery, 2010; Hawkins & Filipović, 2012) related to each level of learners' proficiency.

Methodology

For this study, 200 written productions were selected from the NOCE corpus: 100 samples belonging to first year college students, and another 100 samples belonging to second year college students. All students were enrolled in an English major when the corpus was compiled. The 200 selected samples were randomly taken from the first wave of collection of

the corpus, in years 2003-2004 at the University of Granada. Productions range between 250-300 words and all texts are argumentative, with varied topics, such as the role of new technologies or the importance of learning languages. Although the corpus has been manually annotated with the error tagging system EARS (Díaz-Negrillo, 2009), the samples here used were in raw, unannotated files. Consequently, a pre-processing stage of the texts including the removal of some error tags and lemmatization of the texts was carried out before the computation of readability metrics.

It should be noted that no placement test was taken prior to the compilation of the corpus so that there is no real indicator of learners' proficiency levels. Although students in the first year of their degree are supposed to have a B1 level according to the CEFRL, there is no precise test that confirms this assumption. The same applies to learners in the second year, who are supposed to have a B2 level. As a consequence, the application of readability metrics to see differences between the two groups of learners seems particularly suitable for this corpus, because results may help detecting differences or similarities between the two groups.

Readability metrics. Seventeen measures of readability are going to be used in this study. All the metrics, which are explained below, are available in the R package {koRpus} (Michalke, 2017): ARI, Dale & Chall, Bormuth, DRP, Coleman-Liau, Farr-Jenkins-Paterson, Flesch, Flesch – Kincaid, Fog, FORCAST, Linsear-Write, LIX, RIX, SMOG, Spache, Tuldava, and Wheeler & Smith. For brevity's sake, only a short description of the metrics and their formulae are presented here, readers who wish to know more about the different parameters are referred to the papers where the formulae were originally written.

ARI: The Automatic Readability Index (Senter & Smith, 1967) is computed through the ratios of the number of letters per words, and the number of words per sentence. The original formula gives the result in form of a grade level (GL), but the simplified formula is the one that will be used here:

ARI = (words per sentence) + 9 (word length)

Dale & Chall Readability Formula (Dale & Chall, 1948): it is based on the sentence length and the use of difficult or complicated words. In order to define what a difficult word is, Dale and Chall created a list of 3,000 words. 80% of these words are said to be known by fourth-grade readers. Therefore, words that do not appear in this list denote some difficulty, and they are classified as "hard words". This formula has to be applied to 100 random words selected from a text.

Score: (0.1579 · percentage of hard words)

+ (0.496 · average sentence length) + 3.6365

Bormuth's Mean Cloze and Grade Placement: Bormuth formulae (Bormuth, 1969) are famous because of the use and the prediction and the use of mean clozes. A mean cloze is the percentage of words that have been correctly written in a cloze test. And a cloze test is simply a text where one word, every four or five, has been deleted. The higher the percentage of the mean cloze (therefore, more words have been completed) the easier the text is considered to be. Bormuth's formulae were validated by being contrasted with actual mean cloze tests, instead of being validated by correlating them with other readability metrics.

$$\text{Mean Cloze (M)} = 0.886593 - 0.083640\left(\frac{number\ of\ letters}{number\ of\ words}\right) + 0.161911\left(\frac{number\ of\ words\ from\ the\ Dalle-Chall\ list}{number\ of\ words}\right)^3 - 0.021401\left(\frac{number\ of\ words}{number\ of\ sentences}\right) + 0.000577\left(\frac{number\ of\ words}{number\ of\ sentences}\right)^2 - 0.000005\left(\frac{number\ of\ words}{number\ of\ sentences}\right)^3$$

Degrees of Reading Power (Bormuth, 1969): it consists on the transformation of Bormuth's Mean Cloze on a scale of 0-100, in which 30 means very easy, 50-60 normal difficulty, and 100, very hard. The corresponding formula is:

$$DRP = (1 - \text{Mean Cloze Score}) \cdot 100$$

Coleman-Liau (Coleman & Liau, 1975): the formula first estimates the Cloze Percentage (ECP) and then implements its result to get the final Coleman-Liau score:

$$ECP = 141.8401 - (0.214590 \cdot \text{number of characters}) + (1.079812 \cdot \text{sentences})$$

$$CLI = \left(-27.4004 \cdot \frac{ECP}{100}\right) + 23.06395$$

Flesch Readability Ease (Flesch, 1948): it gives a score in a rank between 0-100, being 0-30 very difficult; 60-70 the standard difficulty, and 90-100 very easy. Flesch's scores are correlated with school grades (USA) and with the estimated percentage of the adult population of the USA (see Flesch, 1949).

$$FRE = 208.835 - (1.015 \cdot \text{average sentence length}) - (84.6 \cdot \text{average number of syllables per word})$$

Farr-Jenkins-Paterson Index (Farr, Jenkins, & Paterson, 1951): it is a recalculation of Flesch index that takes into account the average number of one-syllable words (osw).

$$\text{Score: } 1.599\ \text{osw per 100 words} - (1.015 \cdot \text{average sentence length}) - 31.517$$

Flesch-Kincaid (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975) is used to convert the FRE score into a grade level (GL):

$$GL = (0.39 \cdot \text{average sentence length}) + (11.8 \cdot \text{average number of syllables per word}) - 15.59$$

Fog Index (Gunning, 1952): it is the first index that included the counting of 'hard' words (words with more than two syllables) per 100 words as a variable of the equation to get the readability score:

FOG = 0.4 · (average sentence length + number of 'hard' words)

FORCAST (Caylor, Stitch, & Fox, 1973): originally meant to be used to assess readability in army technical documents, it gives a grade between 5 and 12.9, 12 standing for a really difficult text. The FORCAST rate in a sample of 150 words is calculated in this way:

$$\text{FORCAST} = 20 - \frac{number\ of\ one{-}syllable\ words}{10}$$

Lensear Write Index (O'hayre, 1966): is not actually presented as an index of readability. O'hayre explains that this formula is more about "writeability", that is, it concerns the writer, and not the reader. The formula aims at helping writers to use simple, one-syllable words, that O'hayre (1966, p. 6) considers to be "the words most natural to English". The formula also uses sentence length. If the score is between 70-80, it is an average readable text. Scores above 85 denote simplicity, whereas scores under 65 imply complexity:

Score = number of one-syllable words + (3 · number of sentences)

Björnsson's Läsbarhetsindex, (LIX) was initially designed to be used with texts written in Swedish (Björnsson, 1968). Texts scoring 20 are considered to be very easy, texts scoring 40, average; and text scorings 60 and above, very difficult. The formula was successfully tested when applied to 11 languages (Björnsson, 1983).

LIX = number of words per sentence + percentage of words with more than 6 characters

Anderson's Readability Index, RIX (Anderson, 1981, 1983): it is a modification of the LIX index. RIX was specially designed for English texts, and the score can be easily transformed into a grade level following the chart proposed by Anderson (1983). RIX scores are normally between 1.5 (very easy) and 7.2 or above (very difficult). A text with a RIX score around 3.7 can be considered standard in terms of readability.

$$\text{RIX} = \frac{number\ of\ long\ words}{number\ of\ sentences}$$

Simple Measure of Gobbledygook, SMOG (McLaughlin, 1969): it uses the count of polysyllable words, that is, three or more syllables, in 30 random sentences of the text.

$$\text{SMOG} = 1.0430 \sqrt{number\ of\ polysyllabic\ words} + 3.1291$$

Spache grade (Spache, 1953) was originally designed to analyse readability in textbooks used in primary grades. It uses the average sentence length per 100 words (ASL) and the percentage of "unfamiliar" words. Unfamiliar words are defined as words that do not appear in the list of 769 easy words made by (Dale, 1931).

SG = (0.141 · ASL) + (0.086 · percentage of unfamiliar words) + 0.839

Tuldava Text Difficulty Formula (Tuldava, 1993): It uses the logarithmic transformation, number of sentences, syllables, tokens and words.

$$\text{Tuldava} = \frac{number\ of\ syllables}{number\ of\ tokens} \cdot \ln\left(\frac{number\ of\ words}{number\ of\ sentences}\right)$$

Wheeler & Smith: it was conceived for the use of teachers in primary grades. In their article, the authors offer a table of equivalences of the grades that can be used to interpret the scores (Wheeler & Smith, 1954, p. 398).

$$\text{WS} = \frac{number\ of\ tokens}{number\ of\ sentences} \cdot \frac{10 \cdot \text{number of two-syllable words}}{number\ of\ tokens}$$

As can be seen in this inventory, some metrics are language independent, whereas others are specific to English, Swedish or German. However, the implementation in {koRpus} relies on the English version of TreeTagger (Schmid, 1995), allowing the possibility to extend all the metrics to the investigation of English. It should be noted that language independent methods rely heavily on the number of words and on syllables and mostly try to offer a linear conversion from school grades to complexity scores. After this technical presentation of the metrics, analysis of the scores will be presented in the following section.

Results

First, all productions were POS-tagged using TreeTagger (Schmid, 1995). Second, the 17 readability metrics were computed for both subsets of corpora and the corresponding scores were pooled into a matrix. Third, a last column with the corresponding year of each production was added (year 1, corresponding to students in the first year, and year 2, corresponding to the second year). For all the computed metrics, mean and standard deviations corresponding to the two groups are summarized in Table 1. It is worth noticing that the interpretation of raw results is not straightforward, given that some results are quite counterintuitive. For instance, some of the metrics, such as Bormuth, Coleman.Liau, Farr.Jenkins.Paterson, or Flesch, show that mean scores in year 2 are higher than in year 1, whereas all the other metrics show that mean scores in year 1 are actually higher than in year 2.

In order to see if the differences between the means of the two groups are statistically significant for each metric, a statistical test is needed. Since data was not normally distributed, multiple Kruskal-Wallis tests were performed, as a non-parametric alternative to a t-test. Results are also summed up in Table 1. Only metrics presenting a $p > 0.05$ show a statistically significant difference between the means of year 1 and year 2. That is, four out of the seventeen readability metrics here used (namely Bormuth, Coleman-Liau, DRP, and

FORCAST) do not show statistically significant differences between the means of the two groups. Therefore, one may infer that these metrics do not reflect major variability between the two groups, whereas the other metrics do detect some differences. Yet the interpretability of the results is confusing: even if these four metrics were removed, 13 metrics with their corresponding scores remain to be evaluated. One possible way to spot which metrics are the most accurate in detecting the differences between the two subsets of corpora, and to reduce the number of metrics, is to use machine learning approaches.

Table 1: Descriptive statistics for readability metrics and Kruskal-Wallis results.

	Year 1 (n=100)		**Year 2 (n=100)**		**Kruskal-Wallis test**	
Readability index	**mean**	**SD**	**mean**	**SD**	**X2**	***p***
ARI	10.37	4.50	8.68	2.47	6.8	0*
Bormuth	36.88	4.78	37.73	3.03	2	0.2
Coleman.Liau	54.85	7.74	55.65	4.63	0.6	0.4
Dale.Chall	26.73	8.17	30.18	4.31	7.5	0*
DRP	-3,587	478	-3,673	303	2.1	0.2
Farr.Jenkins.Paterson	63.65	13.61	68.96	6.86	8.5	0*
Flesch	65.67	13.13	71.42	6.41	10.6	0*
Flesch.Kincaid	9.84	3.58	8.23	1.87	10.3	0*
FOG	12.91	3.83	11	2.01	15.2	0*
FORCAST	8.87	0.93	8.68	0.52	1.1	0.3
Linsear.Write	13.75	5.41	11.44	3.01	8.2	0*
LIX	42.24	10.66	36.90	5.16	13.4	0*
RIX	4.46	2.18	3.36	0.89	11.6	0*
SMOG	11.64	2.37	10.62	1.15	11.1	0*
Spache	5.13	1.06	4.74	0.62	5.7	0.02*
Tuldava	4.29	0.60	4.03	0.34	10.3	0*
Wheeler.Smith	60.14	26.37	48.56	12.49	8	0*

Note: *p* values marked with an asterisk are inferior to 0.05 and allow the rejection of the null hypothesis of the Kruskal-Wallis test: equality of ranked means for the two groups.

Support Vector Machines (SVM)

SVM is a type of classifier, a form of machine learning, consisting on algorithms that learn how to detect patterns related to specific classes, or, in this case, particular groups. In this case, a classifier can recognize which of the metrics reflect patterns associated with each one of the four groups; and thus, which of the metrics perform better at classifying productions within groups on the basis of the scores. Here, the use of a classifier does not only account for the assessment of readability metrics, but also for the fact that each group has a particular set of features (i.e. scores associated to each metric) that differentiate it from the other group. Jarvis (2011, p. 130) states that "the classifier-driven approach offers a clearer picture of how well learners' group membership can be predicted on the basis of their language behaviours". Therefore, applying the classifier-driven approach to the present study means to assess how well learners' group membership can be predicted on the basis of their readability metrics scores which are, in a way, a characterization of their language behaviour in written texts.

SVM is a very well-known and powerful discriminative classifier, widely used in supervised learning. It is very popular among the NLP community, in particular in Part-of-Speech (POS) tagging tasks (Giménez & Marquez, 2004), textual semantic similarity (Béchara, et al., 2015), sentiment analysis (Mullen & Collier, 2004; Prabowo & Thelwall, 2009; Song, He, & Fu, 2015), speaker recognition (Campbell, Campbell, Reynolds, Singer, & Torres-Carrasquillo, 2006), or dependency analysis and dependency parsing (Yamada & Matsumoto, 2003; Nivre, Hall, Nilsson, Eryiğit, & Marinov, 2006), among other applications in linguistics.

In the investigation of readability, the use of machine learning (ML) methods has increased in recent years. Since readability metrics, as seen in a previous section of this article, are essentially mathematical formulae, the use of ML is especially suitable for both assessing readability results and the creation of more sophisticated and data-oriented readability metrics. For instance, Shen, Williams, Marius, and Salesky (2013) carried out a study with SVM classifiers in which they showed that the most reliable predictors for text difficulty across languages are length and word-usage features; recurrent parameters of readability formulae. Pilán, Volodina, and Johansson (2014) designed a method for grading sentence readability for leaners of Swedish with SVM incorporated in corpus-based automatically generated exercises through an online platform. Sung, Lin, Dyson, Chang, and Chen (2015) used 30 linguistic features that were fitted into a SVM model to create a readability index for learners of Chinese that automatically classify learners' texts into

CEFRL levels. Sung et al., (2015) showed that SVM methods can be used to improve readability models by adding multilevel linguistic features in a study carried out with a corpus of Chinese textbooks, and Zalmout, Saddiki, and Habash (2016) used readability indices and SVM to show that grammar-based textbooks presented a more coherent progression than communicative-oriented textbooks.

Applying SVM to the data. The choice of SVM among many other machine learning techniques was motivated by several reasons. First, as seen above, there is an increasingly interest for the use of SVM in research on readability metrics, and all readability studies in which SVM has been used have yielded high accuracy results. Second, SVM also appeared to be suitable for the dataset studied here: SVM does not assume data to be normally distributed, SVM algorithms are robust as to the presence of outliers; and most importantly, SVM can deal with collinearity between variables[1]. An implementation of SVM with radial-basis function kernel (RFB) was chosen, being one of the most standard kernels used in scientific studies due to its accuracy (Hsu, Chang, & Lin, 2003; Sung et al., 2015).

The SVM algorithm transposes data into a high dimensional space and creates a hyperplane that allows for the separation and classification of the data. It then selects the subset of the data that is most representative of the dataset and develops, from this subset, the hyperplane that best generalizes for classification. It also calculates a penalization value for the amount of misclassified training data, and this parameter (C) is also included in the formula. Larger values of C may result into more accurate classification, but at the cost of overfitting the data; thus, a model with a lower value of C was preferred in this study.

The goal of applying SVM to the data under scrutiny in this paper is then to predict the groups (year 1, year 2) on the basis of the readability scores. In order to do so, the same data frame, where scores of readability metrics were pooled and class labels (year 1, year 2) were added, was used to compute SVM with the {caret} package. During the training phase, data was split into a training set and a testing set. The training set was used to develop the statistical model that was later applied to make predictions with the testing set. This procedure was repeated 10 times (the data was split into ten subsets, each one of them considered to be independent). As a result, the whole dataset has been part of both the training and the test set, which is known as a ten-fold cross-validation. Eventually, the confusion matrix presented in Table 2 was generated: it shows that most productions belonging to the second year were correctly predicted (except for 9 that were assigned to year 1), but it also shows that the model misclassified 31 year 1 productions, assigned to year 2. In conclusion, it seems that year 2 has some features that differentiate it from year 1, but intrinsic features of year 1 are not so clear-

cut for the model. Accuracy of the model was 80%, and the final values used for the model were sigma = 0.1395 and $C = 1$.

Table 2: Confusion matrix of SVM predictions

n = 200	Year 1 actual	Year 2
Year 1 predicted	69	9
Year 2	31	91

Knowing that the model performed with an accuracy of 80% and that it managed to classify accurately most of the texts belonging to Year 2 – although it struggled with texts from year 1 – the next step was to compute variable importance in order to know which of the 17 metrics were responsible for this classification. This step is crucial, since detecting which metrics allow for classification of the texts will also indicate which are the linguistic parameters that differentiate students from the first and the second year, respectively. Figure 1 shows a plot of the importance of each variable for the model: the three more useful predictors are FOG, LIX, and RIX; whereas the four less important predictors are DRP, FORCAST and Coleman-Liau. Recall that the four less important predictors are, in fact, the four metrics that did not detect any significant difference in the means of the two groups when the Kruskal-Wallis test was performed.

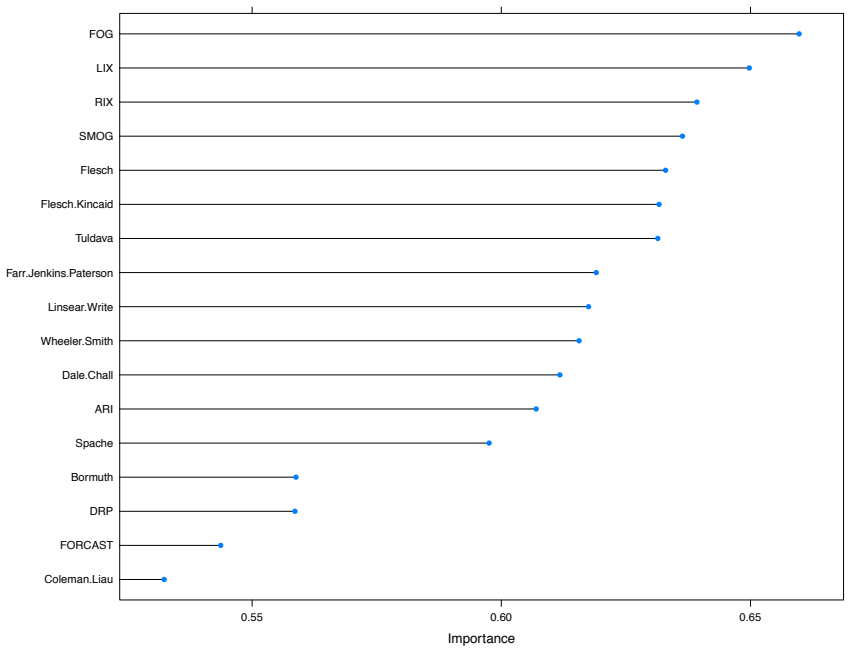

Figure 1: Importance of each readability metric for the SVM model

Conclusions and discussion

Results yielded by the SVM model showed that three metrics were particularly useful to detect differences between the two groups of students: FOG, LIX, and RIX. Linguistic parameters behind these formulae are roughly sentence length, use of polysyllabic words, number of words per sentence, and number of sentences:

FOG = 0.4 · (average sentence length + percentage of polysyllabic words)

LIX = number of words per sentence + percentage of words with more than 6 characters

$$\text{RIX} = \frac{number\ of\ long\ words}{number\ of\ sentences}$$

These three metrics are quite simple to compute, and parameters are almost unmodified by mathematical operations. All in all, it seems that metrics based on very basic aspects of the use of the language, such as sentence length, or the amount of long words, appear to be the most accurate to find differences in the two subsets of corpora under scrutiny. Conversely, metrics with lowest performance at differentiating the groups were those that relied on more complex mathematical operations, and particularly, on the use of mean cloze.

In regards to the complexity detected by the formulae, it should be noted that metrics assume that the use of longer sentences and longer words would result in less easily-readable texts, i.e. more complex productions. However, from the point of view of learners of English, this may not always be the case. One of the intrinsic difficulties of English for Spanish learners is the use of phrasal verbs, which are roughly the combination of a relatively short verb and a preposition (e.g. *put on, check out, get in*….). The (over)use of phrasal verbs may be transposed by some metrics as more simplicity in the average score of readability, whereas Latinate alternatives, often preferred by Spanish learners (e.g. *accelerate*, instead of *speed up*, because of the Spanish counterpart 'acelerar'), are normally longer in syllable length and therefore transposed by the metrics as more complexity in the score. Nevertheless, since Latin-based words are often employed by Spanish learners, this simplistic vision of word length based on approximation of syllable count as indices of complexity may not be accurately applied to Romance language learners of English. Quite the opposite, metrics should rather take into account the use and the variety of phrasal verbs. But because in this paper native speakers and learners have not been compared, and only different groups of learners are taken into consideration, this hypothesis on the use of polysyllabic words has not been validated.

Something similar happens with formulae that include parameters related to the use of words considered to be "difficult", or "sophisticated", on the basis of pre-delimited lists of complex words, such as the Spache, Harris-Jacobson or the Chall-Dale formulae (Spache,

1953; Harris & Jacobson, 1974; Chall & Dale, 1995). Although these metrics have not been particularly useful for the classification of learners in this study, there is an important caveat in their use that should be noted: the implemented word lists were originally designed for native use, and the applicability of these inventories to learners may not be as accurate as expected. Again, consider that learners with Latin L1s may often employ classical or neoclassical compounds, as well as words with Latin origin instead of the Anglo-Saxon equivalents, due to the influence of their L1. Whereas for English native speakers these Greco-Latin words may seem more complicated, or "difficult", for learners with Romance L1s some of these words come up naturally. Conversely, as stated before, the use of phrasal verbs appears to be more complicated to manage. Therefore, a learner-based inventory of "difficult" or complicated words may substantially differ from the one of natives, at least for learners whose L1 is a Romance language. In that sense, a specific index of readability for learners could implement an adapted version of learner wordlist of "easy" or "difficult" words.

All in all, this paper has shown that readability metrics can be used to detect differences in the use of linguistic features among different groups of leaners. Although the two groups compared were assumed to have a different level because learners were in different years of their BA in an English major, no placement test was taken and, therefore, strong claims in terms of the validation of the metrics to detect differences between levels of proficiency cannot be made. This study shows that three readability metrics are particularly useful for the detection of differences in the data under scrutiny – FOG, LIX, and RIX – and confirms that the use of SVM models for classification on the basis of readability-based scores yields good results. Finally, this paper has pointed out that in Second Language Teaching (SLT) and Second Language Learning (SLL), the use of readability metrics and ML techniques is not wide spread. Yet, the use of these techniques would give teachers and SLA researchers new tools to monitor learners' writing progression.

Acknowledgments

I would like to thank my supervisor, Nicolas Ballier, for all his patience, help, and advice. I would also like to thank Taylor Arnold for his explanations on statistics and machine learning, and Ana Díaz-Negrillo, for the compilation of the NOCE corpus. Finally, thanks to Meik Michalke for the creation of the {koRpus} package and for his prompt answers to my doubts through his mailing list.

References

Anderson, J. (1981). Analysing the readability of English and non-English texts in the classroom with Lix. Paper presented at the *Seventh Meeting of the Australian Reading Association*, Darwin, Australia.

Anderson, J. (1983). LIX and RIX: Variations on a little-known readability index. *Journal of Reading*, *26*(6), 490–496.

Béchara, H., Costa, H., Taslimipoor, S., Gupta, R., Orasan, C., Pastor, G. C., & Mitkov, R. (2015). MiniExperts: An SVM approach for measuring semantic textual similarity. (pp. 96–101). *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/s15-2017

Björnsson, C. H. (1968). *Läsbarhet*. Liber.

Björnsson, C.-H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, 480–497. DOI: https://doi.org/10.2307/747382

Bormuth, J. R. (1969). *Development of readability analysis* (Final report project no. 7-0052, contract no. OEC-3-7-070052-0326). US Department of Health, Education and Welfare. Retrieved from https://eric.ed.gov/?id=ED029166

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language*, *20*(2), 210–229. DOI: https://doi.org/10.1016/j.csl.2005.06.003

Caylor, J. S., Stitch, T. G., & Fox. (1973). *Methodologies for determining reading requirements of military occupational specialties.* (Technical Report No. 73-5). Human Resources Research Organization. Retrieved from https://eric.ed.gov/?id=ED074343

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283. DOI: https://doi.org/10.1037/h0076540

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the Association for Information Science and Technology*, *56*(13), 1448–1462. DOI: https://doi.org/10.1002/asi.20243

Dale, E. (1931). A comparison of two word lists. *Educational Research Bulletin*, *10*(18), 484–489.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*(2), 37–54.

Díaz-Negrillo, A. (2007). *A fine-grained error tagger for learner corpora* (Doctoral dissertation). University of Jaen, Jaen.

Díaz-Negrillo, A. (2009). *EARS: A user's manual (vol. 1)*. Munich: LINCOM Academic Reference Books.

DuBay, W. H. (2004). *The principles of readability*. Costa Mesa: Impact Information. Retrieved from http://www.impact-information.com/impactinfo/readability02.pdf

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of Flesch reading ease formula. *Journal of Applied Psychology*, *35*(5), 333–337. DOI: https://doi.org/10.1037/h0062427

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. *Proceedings of the 23rd international conference on computational linguistics* (pp. 276–284). Association for Computational Linguistics.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. DOI: https://doi.org/10.1037/h0057532

Flesch, R. F. (1949). *Art of readable writing*. USA: Hungry Minds Inc.

François, T. (2011). *Les apports du traitement automatique des langues à la lisibilité du français langue étrangère*. Université Catholique de Louvain, Louvain-La-Neuve.

Giménez, J., & Marquez, L. (2004). Fast and accurate part-of-speech tagging: The SVM approach revisited. *Recent Advances in Natural Language Processing III*, 153–162. DOI: https://doi.org/10.1075/cilt.260.17gim

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. DOI: https://doi.org/10.1017/cbo9781139649414

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.

Harris, A. J., & Jacobson, M. D. (1974). *Revised Harris-Jacobson readability formulas.* Paper presented at the Annual Meeting of the College Reading Association, Maryland. Retrieved from https://eric.ed.gov/?id=ED098536

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*(1), 1–23. DOI: https://doi.org/10.1017/s2041536210000103

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework* (Vol. 1). Cambridge: Cambridge University Press.

Heilman, M., Collins-Thompson, K., & Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. *Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 71–79). Association for Computational Linguistics.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Jarvis, S. (2011). Data mining with learner corpora: choosing classifiers for L1 detection. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 127–154). Amsterdam/Philadelphia: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/scl.45.10jar

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. DTIC Document.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. DOI: https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *Tesol Quarterly*, *45*(1) 36–62. DOI: https://doi.org/10.5054/tq.2011.240859

McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading*, *12*(8), 639–646.

McLaughlin, G. H. (1968). Proposals for British readability measures. In J. Downing & A. L. Brown (Eds.), *Third international reading symposium* (pp. 186–205). London: Cassell.

Michalke, M. (2017). Package koRpus: An R Package for Text Analysis (Version 0.10-2). Retrieved from http://reaktanz.de/?c=hacking&s=koRpus

Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. *Proceedings of EMNLP-04, 9th conference on empirical methods in natural language processing* (pp. 412–418). EMNLP.

Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., & Marinov, S. (2006). Labeled pseudo-projective dependency parsing with support vector machines. *Proceedings of the tenth conference on computational natural language learning* (pp. 221–225). Association for Computational Linguistics. DOI: https://doi.org/10.3115/1596276.1596318

O'hayre, J. (1966). *Gobbledygook has gotta go*. US Dept. of the Interior, Bureau of Land Management.

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 174–184). Association for Computational Linguistics. DOI: https://doi.org/10.3115/v1/w14-1821

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. DOI: https://doi.org/10.1016/j.joi.2009.01.003

Schmid, H. (1995). *Treetagger: A language independent part-of-speech tagger* (computer software). Institut Für Maschinelle Sprachverarbeitung: Universität Stuttgart. Retrieved from http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html

Senter, R., & Smith, E. A. (1967). *Automated readability index*. DTIC Document.

Shen, W., Williams, J., Marius, T., & Salesky, E. (2013). *A language-independent approach to automatic text difficulty assessment for second-language learners*. Massachussets Inst. of Technolgy, Lexington Lincoln Lab. DOI: https://doi.org/10.21236/ada595522

Song, J., He, Y., & Fu, G. (2015). *Polarity classification of short product reviews via multiple cluster-based SVM classifiers.* Paper presented at the PACLIC, Shanghai. Retrieved from http://www.aclweb.org/anthology/Y15-2031

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, *53*(7), 410–413. DOI: https://doi.org/10.1086/458513

Sung, Y., Lin, W., Dyson, S. B., Chang, K., & Chen, Y. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, *99*(2), 371–391. DOI: https://doi.org/10.1111/modl.12213

Sung, Y.-T., Chen, J.-L., Cha, J.-H., Tseng, H.-C., Chang, T.-H., & Chang, K.-E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, *47*(2), 340–354. DOI: https://doi.org/10.3758/s13428-014-0459-x

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28*(1), 79–105. DOI: https://doi.org/10.1007/s40593-017-0142-3

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 163–173). Association for Computational Linguistics.

Wheeler, L. R., & Smith, E. H. (1954). A practical readability formula for the classroom teacher in the primary grades. *Elementary English*, *31*(7), 397–399.

Yamada, H., & Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. Proceedings of IWPT (Vol. 3, pp. 195–206).

Zalmout, N., Saddiki, H., & Habash, N. (2016). Analysis of foreign language teaching methods: An automatic readability approach. *Proceedings of the 3rd workshop on natural language processing techniques for educational applications* (122-130). Osaka.

[1] Before running the model, correlations between all pairs of metrics were computed using Spearman 'r', a standard procedure for the correlation of variables in non-normal distributions. This procedure showed that some of the metrics were highly correlated, which may be an issue for some machine learning techniques; and specially, for regression.

Appendix: R script

```
### Investigating the use of readability metrics to detect
differences in written productions of learners: a corpus-based
study. 2017. ###

library(koRpus)
library(tm.plugin.koRpus)

# setting the path for Treetagger
set.kRp.env(TT.cmd="set/your/path/to/treetagger/english",
lang="en",TT.options=list(path="set/your/path/to/treetagger", preset="en"))

# importing the texts individually, but all within a same complex object
noce_year1 <- simpleCorpus(dir=file.path("~/","Desktop","noce_firstyear"),
lang="en",TT.options=list(path="/treetagger/path/here", preset="en"))

## readability---
noce_year1_read <- readability(noce_year1, hyphen = NULL,
  index = c("ARI", "Bormuth", "Coleman.Liau", "Dale.Chall",
    "DRP", "Farr.Jenkins.Paterson",
    "Flesch", "Flesch.Kincaid", "FOG", "FORCAST",
    "Linsear.Write", "LIX","RIX", "SMOG", "Spache",
    "Tuldava", "Wheeler.Smith"), parameters = list(),
```

```
  word.lists    =    list(Bormuth    =    dalechalleasylist,    Dale.Chall    =
dalechalleasylist, Harris.Jacobson = NULL,
    Spache = spachelist), fileEncoding = "UTF-8", tagger = "kRp.env",
  force.lang = NULL, sentc.tag = "sentc", nonword.class = "nonpunct",
  nonword.tag = c(), quiet = FALSE)
corpusSummary(noce_year1_read)
noce_year1_read.df <- corpusSummary(noce_year1_read)
write.table(noce_year1_read.df, file = "noce_year1_ready.csv", sep = ",",
col.names = NA, qmethod = "double")


## SVM---
library(caret)
set.seed(33) # for replicability
x <- subset(all_readability_scores, select=-Year)
y <- Year
# computing the SVM the model
control = trainControl(method="repeatedcv", number=10,
  repeats=3)
model = train(Year~., data=read_melted_years, method="svmRadial",
  preProcess="center", trControl=control)
# predictions of the model
pred <- predict(model,x)
table(pred,y)
# computing and plotting variable importance
importance = varImp(model, scale=FALSE)
plot(importance)
```

Author Information:
Paula Lissón is a graduate student at the University of Paris Diderot (USPC), and currently a visiting student at the department of Linguistics at NYU. She holds an MA in English Linguistics from the University of Paris Diderot (USPC) and a BA in Modern Languages (English and French) from the University of Las Palmas de Gran Canaria, Spain. Her main research interests are psycholinguistics, second and foreign language acquisition, learner corpora, computational linguistics, and natural language processing.
Email: lissonh@gmail.com

To cite this article:
Lissón, P. (2017). Investigating the use of readability metrics to detect differences in written productions of learners: a corpus-based study. *Bellaterra Journal of Teaching & Learning Language & Literature*, *10*(4), 68-86. DOI: https://doi.org/10.5565/rev/jtl3.752